Menselijke machines

INAUGURELE REDE DOOR PROF. DR. MARCEL VAN GERVEN

change perspective

Radboud Universiteit

IN DEI NOMINE FELICITAS

INAUGURELE REDE
PROF. DR. MARCEL VAN GERVEN

Menselijke machines; het lijkt een onmogelijkheid. Wij mensen zijn bewuste levende wezens die kunnen denken, dromen en liefhebben. Machines daarentegen komen over als domme levenloze apparaten die slechts de instructies uitvoeren die wij hen hebben opgedragen. Zal het ooit mogelijk worden om machines te bouwen die daadwerkelijk intelligent gedrag vertonen en zich bewust zijn van hun eigen bestaan? Doorbraken in de artificiële intelligentie (AI) lijken ons steeds dichter bij een positief antwoord op deze vraag te brengen. Maar hoe ver zijn we nu eigenlijk? En hoe kunnen we deze doorbraken gebruiken om de werking van ons eigen brein beter te leren begrijpen? Tijdens deze voordracht zal Marcel van Gerven deze vragen onderzoeken.

Van Gerven onderzoekt de neurale mechanismen van cognitie. Hij bestudeert hoe het brein werkt in zijn natuurlijke omgeving en gebruikt neurale netwerken om menselijke hersenfuncties te modelleren. In toepassingsgericht werk ontwikkelt hij nieuwe machine learning technieken om intelligente machines te ontwikkelen die mensen kunnen evenaren op verscheidene taken. Cognitiewetenschapper Marcel van Gerven was verbonden aan het Max Planck Instituut voor Psycholinguïstiek en het Institute of Ophthalmology, UCL, Londen. Na zijn promotie in de medische informatica heeft hij gewerkt aan brain-computer interfaces en machine learning voor neurale data-analyse. Van Gerven is principal investigator aan het Donders Instituut en hoofd van de vakgroep kunstmatige intelligentie aan de Radboud Universiteit.

Radboud Universiteit

MENSELIJKE MACHINES

Menselijke machines

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar Artificial Cognitive Systems aan de Faculteit der Sociale Wetenschappen van de Radboud Universiteit op vrijdag 14 september 2018

door prof. dr. Marcel van Gerven

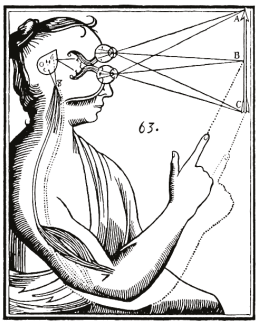
Opmaak en productie: Radboud Universiteit, Facilitair Bedrijf, Post & Print
Fotografie omslag: Bert Beelen

Mijnheer de rector magnificus,
geachte leden van het college van bestuur,
zeer gewaardeerde toehoorders

De titel van mijn voordracht klinkt u misschien wat vreemd in de oren. Menselijke Machines. Het lijkt een *contradictio in terminis*. Wij mensen zijn bewuste, levende wezens die kunnen denken, dromen en liefhebben. Machines daarentegen komen over als domme levenloze apparaten, die enkel de instructies uitvoeren die wij hen hebben opgedragen. Maar is het niet redelijk om de mens als een bijzonder complexe biologische machine te zien? Een machine wiens meningen, verlangens en intenties volledig bepaald worden door de hersenen in samenspel met het lichaam en zijn omgeving? En als we het idee van de mens als machine accepteren, is het dan mogelijk om machines te bouwen die de mens evenaren, of misschien zelfs ooit voorbijstreven? Tijdens deze voordracht vertel ik u wat de huidige stand van zaken is. Vervolgens zullen we bekijken wat ervoor nodig is om menselijkere machines te ontwikkelen, wat deze machines ons kunnen vertellen over onszelf en hoe we deze machines zouden moeten inzetten in onze samenleving. Laten we echter om te beginnen proberen om inzicht te krijgen in het menselijk denken.

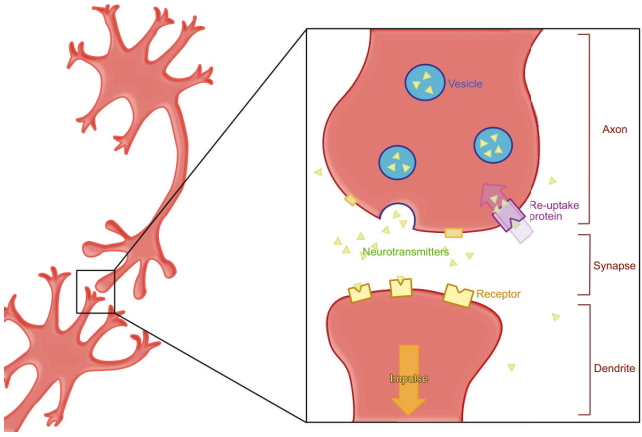
NATUURLIJKE INTELLIGENTIE

De mens is voortdurend in interactie met zijn omgeving. We ontvangen via onze zintuigen een continue stroom van prikkels die ons informatie geven over de wereld om ons heen. Onze waarneming stelt ons dus in staat om onze omgeving te interpreteren. Tegelijkertijd produceren we doorlopend een veelheid aan acties die controle uitoefenen op de buitenwereld. Ons gedrag stelt ons dus in staat om onze omgeving te beïnvloeden[1]. De koppeling tussen waarneming en gedrag wordt tot stand gebracht door onze hersenen. Hierdoor ontstaat een gesloten circuit, dat al in de zeventiende eeuw beschreven is door de Franse filosoof en wiskundige René Descartes[2] (figuur 1).



Figuur 1: De perceptie-actie cyclus volgens Descartes.

Het zijn dus de hersenen die verantwoordelijk zijn voor de denkprocessen die ons handelen bepalen. Laten we eens wat dieper ingaan op hoe complex de menselijke hersenen eigenlijk zijn. Onze hersenen bestaan uit zo'n 86 miljard hersencellen[3], ook wel neuronen genaamd (figuur 2). Deze neuronen communiceren met elkaar door middel van het versturen van boodschappen, ook wel actiepotentialen genoemd. Op het moment dat een actiepotentiaal op de plaats van bestemming aankomt, ook wel een synaps genoemd, wordt hij vertaald in een chemisch signaal dat opgepikt kan worden door het ontvangende neuron. Het gehele brein bevat zo'n 100 biljoen van deze synapsen. Veranderingen in deze synaptische verbindingen stellen het brein in staat om te leren. Hierdoor kunnen we adaptief reageren op de complexe en onzekere fysische, biologische en culturele processen in de wereld om ons heen. Dit is wat ik versta onder *natuurlijke intelligentie.* Het brein kan dus gezien worden als een orgaan dat ons bevrijdt van puur reflexmatig gedrag.



Figuur 2: Neuronen en synapsen.

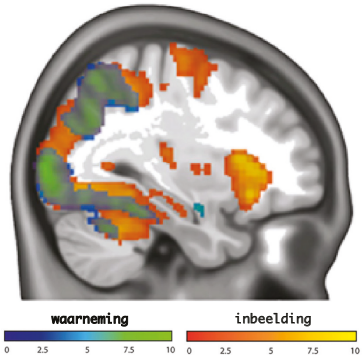
Maar hoe kunnen we inzicht krijgen in de aard van ons denken? Het antwoord hierop hangt af van aan wie men de vraag stelt. De behavioristen uit het begin van de vorige eeuw stelden dat het antwoord op deze vraag buiten ons bereik ligt, aangezien we geen objectieve toegang hebben tot onze eigen denkprocessen. Dit perspectief veranderde echter in de loop van de twintigste eeuw dankzij de opkomst van de cognitiewetenschap, waarin het onderzoek naar het denken, ofwel cognitie, juist centraal staat. Deze cognitieve revolutie werd gevoed door doorbraken binnen zowel de cognitieve psychologie als de artificiële intelligentie.

COGNITIEVE PSYCHOLOGIE

Een grensverleggende ontwikkeling binnen de cognitieve psychologie was het meetbaar maken van mentale processen. Dé pionier op dit gebied was Franciscus Cornelis Donders, die exact tweehonderd jaar geleden werd geboren en naamgever is van het Donders Instituut. Met behulp van een noëmatachograaf vergeleek Donders de reactietijden van proefpersonen onder verschillende experimentele condities (figuur 3). Hierdoor was hij in staat om vast te stellen hoeveel tijd een specifiek denkproces kost[4]. Deze aanpak, die bekend staat als de subtractiemethode, geeft dus op basis van *extern* observeerbaar gedrag (reactietijden) toegang tot *interne* mentale toestanden die anders voor ons verborgen zouden blijven.

Figuur 3: Donders en zijn noëmatachograaf[5].

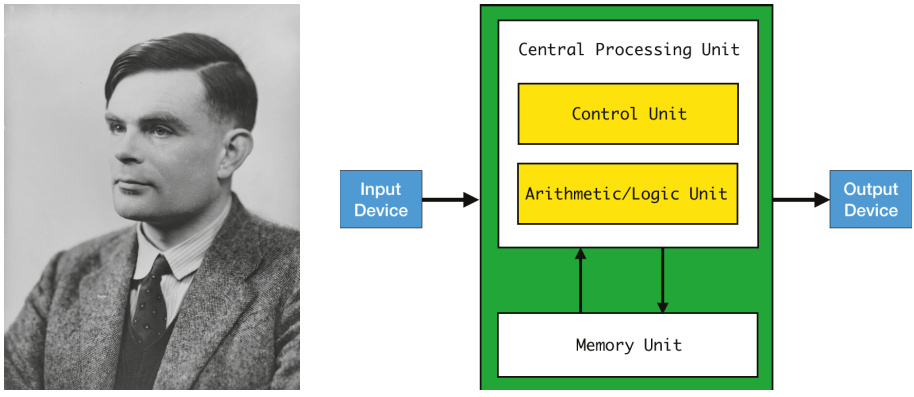
Het werk van Donders is van grote invloed geweest op de cognitieve neurowetenschap. Dit vakgebied onderzoekt hoe we het menselijk denken kunnen relateren aan specifieke hersenprocessen. Waar men eerder aangewezen was op het begrijpen van hersenfuncties via onderzoek naar patiënten met een hersentrauma, hebben onderzoekers tegenwoordig de beschikking over technieken als *magnetic resonance imaging*, ofwel MRI, om hersenactiviteit in kaart te brengen. Ook hier wordt veelvuldig gebruik gemaakt van Donders' subtractiemethode om te achterhalen welke hersengebieden actief worden bij het uitvoeren van een taak. De kleurrijke MRI-scans die u vaak in de media ziet, zijn het resultaat van deze analyse. Hieronder ziet u een voorbeeld uit ons eigen werk[6]. We tonen hier welke hersengebieden actief worden wanneer we een object *waarnemen* versus wanneer we ons datzelfde object *inbeelden*. Een interessante observatie is dat overlappende hersengebieden actief worden. Dit suggereert dat dezelfde hersenprocessen een rol spelen bij zowel waarneming als inbeelding.



Figuur 4: De relatie tussen waarneming en inbeelding.

ARTIFICIËLE INTELLIGENTIE

Ook de ontwikkeling van de artificiële intelligentie, ofwel AI, heeft grote invloed gehad op het denken over ons denken. Het idee van een intelligente machine gaat terug tot de Griekse oudheid en is intensief bestudeerd tijdens de Verlichting[7]. Het daadwerkelijk bouwen van intelligente machines moest wachten op de komst van de computer. De Britse wiskundige Alan Turing formaliseerde in de jaren dertig van de vorige eeuw de notie van berekenbaarheid[8]. Hij toonde aan dat een hypothetische universele machine alle problemen kan oplossen die mechanisch berekenbaar zijn. Het recept voor de oplossing komt in de vorm van een *algoritme*; een reeks instructies die uit te voeren zijn door een computer. Turings werk vormt de inspiratie voor de Von Neumann-architectuur, die ten grondslag ligt aan de moderne digitale computer[9] (figuur 5).



Figuur 5: Alan Turing en de Von Neumann-architectuur[10].

De cognitiewetenschap gaat ervan uit dat ook onze eigen mentale toestanden mechanisch berekenbaar zijn en fysisch gerealiseerd worden door ons brein. Onze geest ligt dus besloten in de voortdurend veranderende patronen van hersenactiviteit die betrokken zijn bij neurale informatieverwerking. Dit alles heeft een cruciale implicatie: *Intelligentie kan in principe worden nagebootst met een computer*. Deze stelling is niet alleen fundamenteel voor de AI, maar ook essentieel voor de cognitiewetenschap[11]. Het bouwen van menselijke machines is namelijk noodzakelijk om onze theorieën over natuurlijke intelligentie te kunnen toetsen[12]. Om met de woorden van Richard Feynman te spreken: *What I cannot create, I do not understand*.

DE FORMALISERING VAN RATIONALITEIT

Om deze machines te kunnen bouwen, is het van belang om helder te krijgen wat we eigenlijk bedoelen met intelligentie. Er zijn hier drie ingrediënten van belang (figuur 6). Ten eerste dient een intelligent wezen in staat te zijn tot *actie*, aangezien hiermee gedrag tot stand gebracht kan worden[13]. Ten tweede dienen we uit te kunnen drukken hoe wenselijk de situatie is die ontstaat als we een actie hebben uitgevoerd. Deze wenselijkheid wordt ook wel *utiliteit* genoemd. Om intelligentie te begrijpen, moeten we dus weten welke doelen een intelligent wezen nastreeft[14]. Ten derde moeten we bij het nemen van een beslissing altijd onze *onzekerheid* over de toestand van de wereld meenemen. Deze onzekerheid komt voort uit de ambiguïteit van onze waarneming, onze beperkte kennis, en de onvoorspelbare gevolgen van ons handelen[15]. We kunnen nu het volgende stellen: *Een wezen gedraagt zich intelligent als het dié actie selecteert die de verwachte utiliteit maximaliseert*[16].

Figuur 6: Actie, utiliteit en onzekerheid als ingrediënten van intelligentie.

Een belangrijk gevolg van deze stelling is dat we moeten kunnen redeneren onder onzekerheid. Dit is echter niet eenvoudig. Het vereist namelijk dat we alle mogelijke toestanden van de wereld in acht nemen. Iedere toestand heeft immers een (wellicht minieme) kans om de werkelijke toestand van de wereld te zijn. Om deze kansen uit te rekenen, kunnen we gebruik maken van een wiskundige formule die geïntroduceerd is door de 18e-eeuwse predikant Thomas Bayes[17]. De regel van Bayes beschrijft hoe onze onzekere kennis van de wereld, vastgelegd in termen van een kansmodel, aangepast

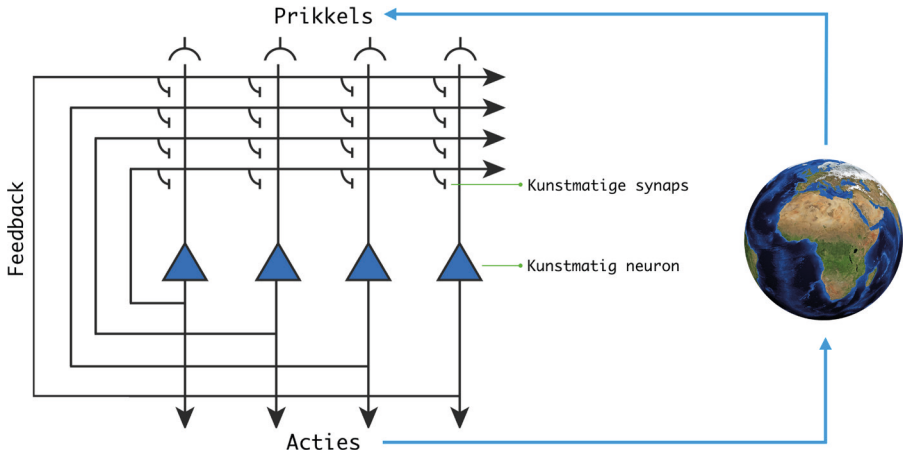
moet worden op basis van onze waarneming[18]. Exact redeneren onder onzekerheid kost echter onredelijk veel rekenkracht. Er is dan ook door ons en anderen veel aandacht besteed aan de ontwikkeling van methoden die exact redeneren optimaal benaderen[19]. De Bayesiaanse statistiek is van groot belang geweest voor de ontwikkeling van intelligente systemen. Zo heb ik in mijn eigen onderzoek modellen ontwikkeld binnen de oncologie om de effecten van chemotherapie op de kwaliteit van leven te voorspellen[20]. Het Bayesiaanse paradigma speelt echter ook een cruciale rol in de formulering van moderne theorieën over natuurlijke intelligentie. De zogenaamde Bayesiaanse breinhypothese stelt dat ons brein zich gedraagt alsof het exact redeneren benadert[21]. Ze gaat ervan uit dat het brein een intern model van de realiteit opbouwt op basis van onze waarneming en ons handelen. Onder deze interpretatie leven we als het ware in een *gecontroleerde hallucinatie,* waarin onze verwachtingen voortdurend worden bijgestuurd door onze sensaties en worden getoetst door onze acties[22]. Dat onze waarneming sterk gekleurd wordt door onze verwachting blijkt bijvoorbeeld uit de holle masker illusie[23]. We kunnen de binnenkant van het masker niet als hol zien door de dominante invloed van onze voorkennis.

Samenvattend biedt het Bayesiaanse gedachtengoed een elegant normatief kader waarmee we zowel menselijk als kunstmatig redeneren kunnen beschrijven. Het vertelt ons echter niet hoe een wezen leert om zich binnen dit kader te gedragen.

CONNECTIONISME

De Bayesiaanse aanpak geeft een abstracte *top-down* beschrijving van intelligentie. Een andere manier om intelligentie te bestuderen is door ons juist te richten op de elementaire bouwstenen van informatieverwerking in ons brein en de neurale mechanismen die daar plaatsvinden na te bootsen. Deze biologisch geïnspireerde aanpak gaat terug naar de begindagen van de digitale computer en leidde tot de opkomst van neurale netwerken als modellen van neurale informatieverwerking[24]. Een neuraal netwerk bestaat uit een groot aantal kunstmatige neuronen die met elkaar verbonden zijn middels kunstmatige synapsen (figuur 7). Door het netwerk te verbinden met zijn omgeving kan het acties uitvoeren op basis van binnenkomende prikkels. Door feedback te introduceren krijgt het neurale netwerk de mogelijkheid om te reflecteren op het verleden en te speculeren over de toekomst.

Bij het bouwen van een neuraal netwerk kunnen we ons tot doel stellen om biologisch zo realistisch mogelijke modellen te maken[25]. Deze aanpak staat aan de basis van de computationele neurowetenschap. Stel nu dat we de toestand van uw eigen brein zeer precies zouden kunnen meten. We zouden dan met behulp van zeer gedetailleerde neuronmodellen in theorie een synthetisch brein kunnen bouwen wiens gedachten niet te onderscheiden zijn van uw eigen gedachten. Met andere woorden, we zouden een directe kopie van uw geest kunnen creëren. Naast de ethische bezwaren die opgeworpen kunnen worden, is het op deze wijze nabootsen van de hersenen praktisch ge-



Figuur 7: Een neuraal netwerk.

zien een onmogelijke opgave. Het vereist ten eerste dat onze wiskundige vergelijkingen compleet zijn. Dit is bij lange na niet het geval. Ten tweede vereist het dat we het brein perfect in kaart kunnen brengen. Ondanks grote doorbraken in de neurotechnologie zijn we hier nog ver van verwijderd. Ten derde veronderstelt het dat onze computers genoeg capaciteit hebben om een realistische simulatie van het brein te kunnen draaien. Ondanks steeds krachtigere hardware is ook dit nog verre van haalbaar. Dat het niet triviaal is om een brein te simuleren, blijkt ook uit het OpenWorm project (http://openworm.org). Het doel van dit project is om het gedrag van de worm C. Elegans na te bootsen in een computer. Dit blijkt echter verre van triviaal te zijn. Ter overpeinzing: het zenuwstelsel van deze worm bestaat uit slechts 302 neuronen.

Een andere werkwijze, die gebruikt wordt in de AI, is om de biologische details juist te negeren en enkel de essentie van neurale informatieverwerking te behouden. Dit resulteert in eenvoudigere en daardoor beter hanteerbare modellen[26]. In deze modellen wordt de sterkte van iedere synaptische verbinding vastgelegd door middel van een getal, ook wel gewicht genaamd. Het gedrag van een neuraal netwerk hangt uiteindelijk af van de waardes van deze gewichten, net zoals het gedrag van ons brein (grotendeels) afhangt van de verbindingen tussen neuronen.

We kunnen een neuraal netwerk gebruiken als het synthetische brein van een machine. De grote vraag is hoe we het netwerk zodanig kunnen instellen dat het een specifiek probleem kan oplossen. Hiertoe zijn leerregels zoals het populaire *backpropagation* algoritme ontwikkeld. Dit algoritme vertelt ons hoe de gewichten van een neuraal netwerk aangepast moeten worden aan de hand van data om zo een specifiek doel te realiseren[27]. Een neuraal netwerk is dus een voorbeeld van een *zelflerend systeem*. We kunnen drie varianten van leren onderscheiden, afhankelijk van de gegevens die er

voorhanden zijn. Bij *supervised learning* nemen we aan dat het juiste antwoord beschikbaar is tijdens het leren. Denk bijvoorbeeld aan een ouder die haar kind antwoord geeft op een vraag. Bij *reinforcement learning* is er enkel een feedbacksignaal aanwezig aan de hand waarvan het netwerk zijn gedrag kan bijstellen. Denk bijvoorbeeld aan de verandering van uw gedrag als u zich brandt aan een fornuis. Bij *unsupervised learning* heeft het netwerk enkel beschikking over zijn sensaties op basis waarvan het leert om patronen te ontdekken. Denk bijvoorbeeld aan een baby die leert om haar omgeving te begrijpen.

Neurale netwerken staan ook aan de basis van het *connectionisme*. Dit is de stroming binnen de cognitiewetenschap die neurale netwerken gebruikt om menselijke cognitie te bestuderen[28]. Connectionisten benadrukken de belangrijke overeenkomsten die neurale netwerken vertonen met ons eigen brein. Zo maken neurale netwerken, net zoals onze hersenen, gebruik van parallelle gedistribueerde informatieverwerking. Dit houdt in dat informatie simultaan door het hele netwerk heen verwerkt wordt. Dit is in tegenstelling tot de sequentiële gecentraliseerde informatieverwerking in digitale computers[29]. Ook het leergedrag van neurale netwerken en hun aanpassingsvermogen bij schade vertonen een interessante gelijkenis met de plasticiteit en weerbaarheid van ons eigen brein. Het connectionisme biedt dus een aantrekkelijk raamwerk waarbinnen we kunnen onderzoeken hoe cognitie kan ontstaan vanuit de interactie tussen eenvoudige componenten. Het geeft ons een bottom-up aanpak, waarmee we intelligent gedrag kunnen nabootsen in een machine. Laten we eens kijken waar deze intelligente machines in de praktijk toe in staat zijn.

MACHINES IN ONS MIDDEN

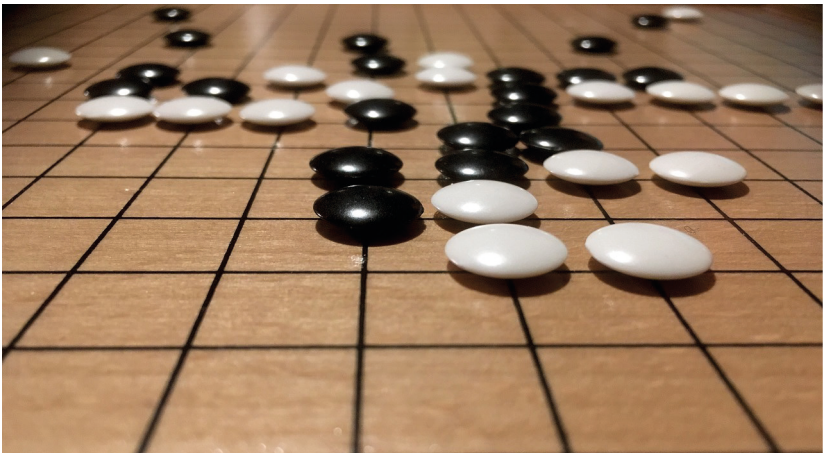
Zoals al genoemd, is een van de karakteristieke eigenschappen van intelligentie het kunnen interpreteren van onze omgeving. Denk bijvoorbeeld aan het evolutionaire voordeel dat onze voorouders hebben gehad aan het kunnen onderscheiden van roofdieren en prooidieren. Decennialang was het onmogelijk om computers een vorm van waarneming te geven die ook maar enigszins in de buurt kwam van die van de mens. Enkele jaren geleden kwam computer-gebaseerde waarneming echter binnen handbereik door de opkomst van *deep learning*[30]. Dit verwijst naar het trainen van neurale netwerken die uit een groot aantal opeenvolgende lagen van kunstmatige neuronen bestaan[31]. Iedere laag leert hier om steeds complexere eigenschappen van de ingevoerde stimulus te representeren. De toepassing van diepe neurale netwerken leidde tot een doorbraak op het gebied van de beeldverwerking. Eindelijk was het mogelijk om computers willekeurige objecten te laten herkennen. Momenteel zelfs in die mate dat ze bovenmenselijk goed presteren op deze taak.

Deep learning heeft tot verschillende toepassingen geleid die tot voor kort onmogelijk waren. Binnen de geneeskunde zijn neurale netwerken nu bijvoorbeeld vaak beter in het herkennen van afwijkingen dan artsen met jarenlange ervaring. Neurale netwer-

ken worden ook gebruikt bij de ontwikkeling van zelfrijdende auto's. Door het camerabeeld te segmenteren in betekenisvolle onderdelen, kan de zelfrijdende auto zijn omgeving beter interpreteren. Zelf hebben wij een neuraal netwerk ontwikkeld dat in staat is om schetsen om te zetten in fotorealistische afbeeldingen[32]. De meest recente variant werkt zelfs voor willekeurige fictieve personages. Wat dacht u bijvoorbeeld van de onderstaande interpretaties van respectievelijk Barbie, de Mona Lisa en Sneeuwwitje (figuur 8)? Ook persoonlijke digitale assistenten gebruiken neurale netwerken voor het herkennen van spraak, het vertalen van tekst, en het beantwoorden van vragen. Als u gebruik maakt van een computerprogramma als Siri of Alexa dan bent u dus in feite in gesprek met een neuraal netwerk.

Figuur 8: Barbie, de Mona Lisa en Sneeuwwitje volgens een neuraal netwerk.

Tot dusverre hebben wij ons gericht op machines die goed zijn in het interpreteren van hun omgeving. Intelligente machines moeten echter ook in staat zijn om complexe problemen op te lossen. Hiervoor moeten ze de juiste beslissing op het juiste moment kunnen nemen. Computerspellen hebben al sinds de begindagen van de AI een belangrijke rol gespeeld om het probleemoplossend vermogen van intelligente machines te testen. Een van de eerste spellen die tegen de computer gespeeld kon worden was boter, kaas en eieren. Dit computerspel was geïmplementeerd op de EDSAC; een van de eerste naoorlogse computers[33]. De heilige graal binnen de AI is echter het winnen van het spel Go (figuur 9). Het aantal manieren waarop dit spel kan verlopen, is namelijk groter dan het aantal atomen in ons heelal[34]. Ook hier hebben recente ontwikkelingen voor een doorbraak gezorgd. Onderzoekers trainden diepe neurale netwerken met behulp van *reinforcement learning* om zowel de waardering van iedere bordpositie als de kans op een volgende zet te berekenen. Door deze netwerken te combineren met een geavanceerde zoekstrategie leerde de computer om zelfs de beste spelers ter wereld met gemak te verslaan[35].



Figuur 9: Het spel Go.

Go is echter nog relatief eenvoudig. De bordpositie is volledig observeerbaar, de regels van het spel zijn bekend, per zet hebben we een handvol mogelijkheden, de uitkomst van een gekozen zet ligt vast en het spel heeft een beperkte duur. Dit staat in schril contrast met de complexe situaties waarmee u en ik in de dagelijkse realiteit worden geconfronteerd. De wereld om ons heen is namelijk slechts deels observeerbaar, het aantal mogelijke waarnemingen en handelingen is bijna oneindig groot, de gevolgen van een handeling zijn onzeker en de consequenties van een actie kunnen zich pas ver in de toekomst manifesteren. Dezelfde ingrediënten vinden we terug in computerspellen zoals DOTA 2, waarin een virtueel wezen moet zien te overleven in een virtuele wereld. Waar het winnen van dit soort games tot voor kort een brug te ver leek, zijn ook hier recentelijk doorbraken behaald door de ontwikkeling van nieuwe leeralgoritmen[36].

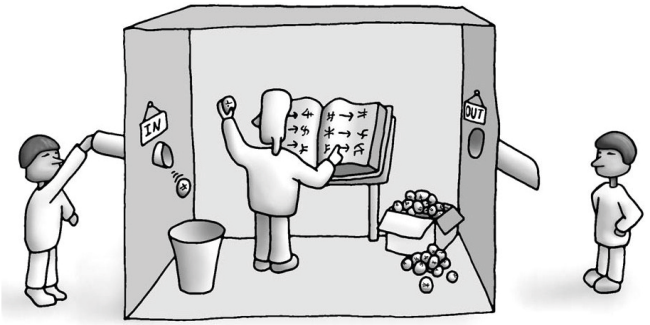
ARTIFICIËLE STUPIDITEIT

Missie geslaagd zou u denken. Als we de nieuwste generatie robots voorzien van slimme algoritmen dan lijken menselijke machines om de hoek te staan. Toch vertonen de huidige intelligente machines vooralsnog een hoge mate van artificiële stupiditeit.

Een voorbeeld is hoe de perceptie van een neuraal netwerk eenvoudig verstoord wordt. Zo kan het camerabeeld van een zelfrijdende auto door het toevoegen van een minieme hoeveelheid ruis plotseling geen voetgangers meer herkennen[37]. We noemen dit ook wel een *adversarial example*. U kunt zelf wel bedenken wat de dramatische consequenties van dit soort fouten kunnen zijn.

Ook bij het leren handelen in complexe omgevingen lopen we tegen beperkingen aan. In het geval van DOTA 2 leert de computer door tegen zichzelf te spelen gedurende een periode die gelijk staat aan 180 mensenjaren. De computer heeft dus extreem veel voorbeelden nodig voordat hij enig zinvol gedrag vertoont. Dit in tegenstelling tot mensen en dieren, die zelfs op basis van één enkel voorbeeld kunnen leren om radicaal ander gedrag te vertonen. Het generaliseren naar nieuwe situaties blijkt ook erg lastig te zijn. Het kunnen spelen van DOTA biedt de computer bijvoorbeeld weinig tot geen voordeel om een ander spel zoals StarCraft te leren spelen. Verder is het nog maar de vraag in hoeverre resultaten behaald in virtuele werelden relevant zijn voor het effectief leren handelen in de echte wereld[38]. Het beperkte succes van volledig zelfstandig handelende robots is hier een mooi voorbeeld van.

Hoe komt het dat intelligente machines nog steeds zo breekbaar zijn? Het komt er in grote lijnen op neer dat ze in staat zijn om complexe problemen op te lossen zonder daadwerkelijk besef te hebben van zichzelf of de wereld om hen heen. De tot dusverre beschreven neurale netwerken kunnen we als het ware interpreteren als *geavanceerde behavioristische reflexmachines*. Dit doet ons denken aan Searle's Chinese kamer[39] (figuur 10). In dit gedachtenexperiment beantwoordt iemand in een kamer vragen die in het Chinees worden gesteld. Door met behulp van een instructieboek het bijbehorende Chinese antwoord op te zoeken, lijkt het voor de mensen buiten de kamer alsof de persoon in de kamer Chinees spreekt. In werkelijkheid is er natuurlijk geen enkel begrip van de Chinese taal. Net zo hebben hedendaagse intelligente machines geen flauw benul van de taak waar ze mee bezig zijn. Ze missen de *intentionaliteit* en *flexibiliteit* die zo kenmerkend zijn voor natuurlijke intelligentie.



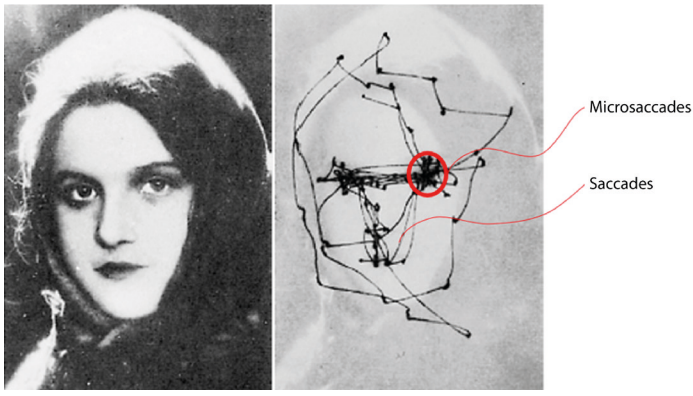
Figuur 10: Searle's Chinese kamer.

DE MENS IN DE MACHINE

De grote vraag blijft dus of we menselijke machines kunnen bouwen die niet alleen enorm goede reflexen hebben, maar ook net zoals wij kunnen overleven in een complexe en veranderlijke wereld. Een sterkere integratie van wiskundige technieken uit de

artificiële intelligentie met inzichten uit de cognitieve psychologie en de neurowetenschap biedt perspectieven.

Vanuit empirisch oogpunt kunnen we proberen om synthetische breinen te bouwen die op dezelfde manier reageren als hun biologische tegenhangers. De cognitieve psychologie kan bijvoorbeeld inzichten over leren en gedrag bieden om zo menselijkere machines te bouwen. Een specifiek voorbeeld is het simuleren van oogbewegingen (figuur 11). Als wij een stimulus waarnemen dan zullen we actief onze ogen naar die plekken bewegen die ons met zo min mogelijk moeite zoveel mogelijk informatie over de stimulus geven[40]. Door neurale netwerken te ontwikkelen die de oogbewegingen van mensen nabootsen, kunnen we machines bouwen die niet alleen natuurgetrouwer, maar ook efficiënter werken[41]. Om menselijke machines te bouwen, zullen we ons ook moeten verdiepen in de vraag welke utiliteit wijzelf proberen te maximaliseren. Zoals ik eerder betoogde, is het deze utiliteit die uiteindelijk ons gedrag bepaalt. De utiliteitsfunctie van ons mensen is echter uitermate complex en aan verandering onderhevig. Ze moet namelijk niet alleen uitdrukking geven aan onze primaire driften, maar ook aan de verscheidenheid aan intrinsieke en extrinsieke motivaties die ons gedrag bepalen. De ontwikkeling van machines die vergelijkbare doelen nastreven in een virtuele of echte wereld is dan ook essentieel om daadwerkelijk intelligent gedrag te bewerkstelligen.

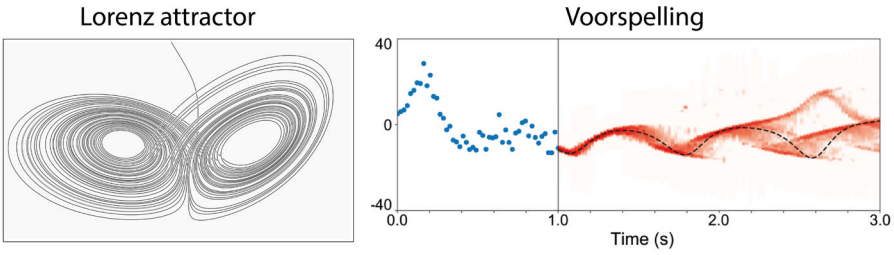


Figuur 11: Een stimulus en zijn geassocieerde oogbewegingen[42].

Ook de neurowetenschap speelt een belangrijke rol in het bouwen van menselijke machines. Een overtuigend voorbeeld is hoe diepe neurale netwerken geïnspireerd zijn op klassiek neurowetenschappelijk onderzoek naar visuele waarneming[43]. Voortschrijdend onderzoek naar neurale informatieverwerking kan wellicht nieuwe (mechanistische of functionele) principes ontdekken die overdraagbaar zijn naar machines[44]. We zien in dit kader ook een herwaardering van brein-geïnspireerde (*neuromorphic*) hard-

ware[45]. Deze hardware kan vele malen efficiënter opereren dan die van hedendaagse computers en zal een belangrijke rol gaan spelen in de ontwikkeling van toekomstige intelligente machines.

Vanuit theoretisch oogpunt zien we dat er een verdere verfijning en integratie van bestaande technieken plaatsvindt. Ik zie de Bayesiaanse en connectionistische paradigma's dan ook als twee kanten van dezelfde medaille[46]. Het Bayesiaanse paradigma geeft ons een top-down aanpak die formaliseert hoe intelligente machines zich zouden moeten gedragen. Het connectionisme geeft ons een bottom-up aanpak, waarmee dit optimale gedrag benaderd kan worden. Zo hebben wij bijvoorbeeld in recent werk laten zien dat onzekere situaties erg goed te benaderen zijn met behulp van neurale netwerken. Dit stelt ons in staat om zeer precieze voorspellingen te maken van de toekomstige toestand van complexe systemen (figuur 12). Dit sluit naadloos aan bij het idee dat ook ons brein een voorspellingsmachine is[47].



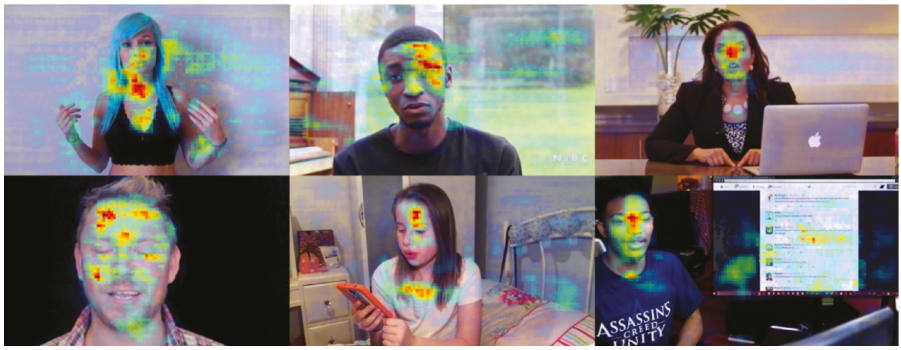
Figuur 12: Voorspelling van de evolutie van een Lorenz attractor.

Eerder heb ik benadrukt dat intelligente wezens over een intern model van de realiteit dienen te beschikken, waarmee ze de consequenties van hun handelen kunnen voorspellen. Misschien kunnen we machines forceren om een intern model te leren, waarmee ze hun eigen toekomst kunnen voorspellen. Recent onderzoek laat zien dat we inderdaad neurale netwerken kunnen trainen om te voorspellen tot welke toekomstige situaties hun acties zullen leiden[48]. Het denken kunnen we in deze zin interpreteren als *gesimuleerd gedrag*[49]. Om te kunnen denken legt het neurale netwerk een intern model vast in zijn synaptische verbindingen. De ontwikkeling van zo'n intern model is een minimale voorwaarde voor het ontstaan van intentionaliteit. Op het moment dat een machine leert om zijn eigen toekomstige gedrag te voorspellen, ontstaat een rudimentair zelfbeeld. Door het toekomstige gedrag van andere wezens te leren voorspellen ontstaat een besef van de ander, en daarmee ook de mogelijkheid tot communicatie. Ik begeef mij nu op glad ijs, maar wellicht is bewustzijn een noodzakelijke eigenschap van een machine die in staat is om al haar sensaties te verklaren[50].

DE MACHINE IN DE MENS

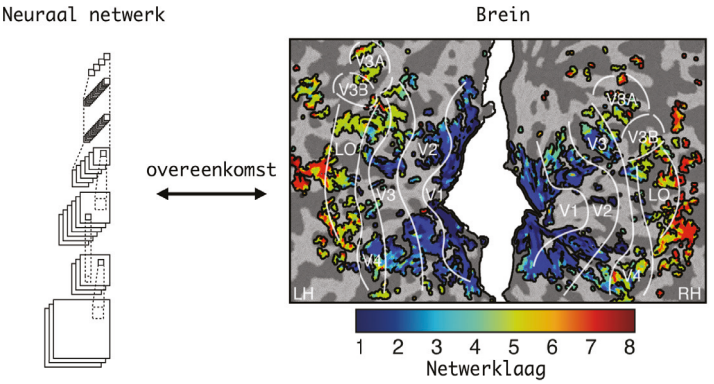
We kunnen dus theoretische ontwikkelingen en neuropsychologische inzichten gebruiken om menselijkere machines te bouwen. Maar hoe kan intelligente technologie ons helpen om onszelf beter te begrijpen?

De cognitieve psychologie bestudeert menselijk leren en gedrag. Als we een synthetisch brein kunnen bouwen dat zich hetzelfde gedraagt als wijzelf, dan kunnen we dit brein bevragen om inzicht in ons eigen gedrag te krijgen. Zo hebben wij bijvoorbeeld een neuraal netwerk getraind om iemands (waargenomen) persoonlijkheidskenmerken te voorspellen (figuur 13). Door de interne toestanden van het netwerk te analyseren, krijgen we inzicht in hoe wij tot vergelijkbare voorspellingen komen. We kunnen hiermee de vooroordelen en stereotyperingen blootleggen die ons gedrag kunnen bepalen. Ander recent onderzoek in ons lab richt zich op het automatisch beschrijven van het gedrag van mens en dier. Dit maakt het eenvoudiger om cognitie in het wild te bestuderen.



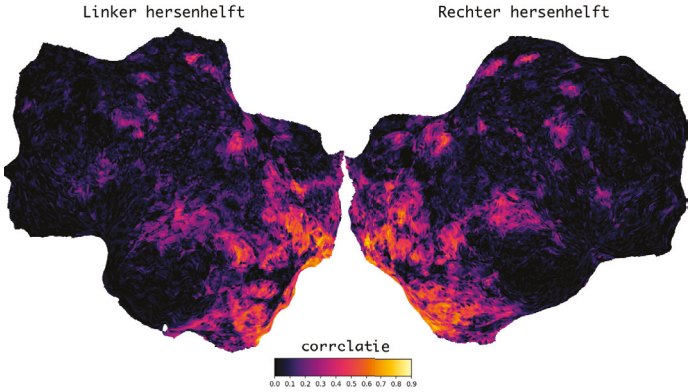
Figuur 13: Visualisatie van beeldelementen die een neuraal netwerk (waargenomen) persoonlijkheidskenmerken laat voorspellen.

De neurowetenschap heeft op haar beurt als doelstelling om neurale informatieverwerking te begrijpen. Mijn stelling is dat we bepaalde eigenschappen van ons brein kunnen verklaren door synthetische breinen bloot te stellen aan die problemen waarmee ook wij geconfronteerd worden. Zo hebben wij bijvoorbeeld laten zien dat er een interessante overeenkomst bestaat tussen kunstmatige en biologische neurale netwerken (figuur 14). Het blijkt dat de interne representaties van diepe neurale netwerken, die getraind zijn om objecten te categoriseren, een gelijkenis vertonen met neurale representaties in ons brein[51].



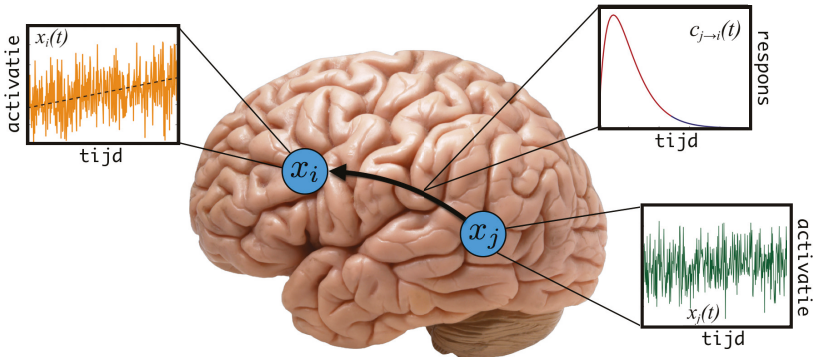
Figuur 14: Overeenkomst tussen kunstmatige en biologische neurale netwerken.

Ik zie dit als een belangrijke stap op weg naar een computationele benadering van de cognitieve neurowetenschap. De conventionele aanpak is om te achterhalen waar en wanneer het gemiddelde brein actief is onder een specifieke experimentele manipulatie. Dit vertelt ons echter niets over de neurale informatieverwerking die hieraan ten grondslag ligt. Een informatievere aanpak is om synthetische breinen te ontwikkelen die (mechanistische of functionele) verklaringen bieden voor de gedragsmatige en neurale observaties van *individuele* organismen in hun *natuurlijke* omgeving. Ter illustratie ziet u in figuur 15 hoe we met behulp van een eenvoudig synthetisch brein de neurale activiteit van een van onze Masterstudenten voorspellen, nadat we hem 24 uur lang in een MRI-scanner naar de serie Doctor Who hebben laten kijken[52]. De correlatie tussen het synthetische brein en het echte brein is vooral in visuele hersengebieden zeer hoog.



Figuur 15: Voorspelling van hersenactiviteit tijdens het kijken naar de serie Doctor Who.

Meer in het algemeen zal de inzet van intelligente technologie steeds crucialer worden bij het verder ontrafelen van onze hersenen, zeker gegeven de steeds grotere neurale datasets die verzameld worden. Zo hebben wij technieken ontwikkeld om de anatomische structuur van ons brein te voorspellen[53], de functionele organisatie van ons brein beter in kaart te brengen[54] en de causale interacties tussen neuronen te kwantificeren[55]. Van het laatste ziet u in figuur 16 een voorbeeld. Op termijn kan AI niet alleen nieuwe inzichten bieden in de neurale mechanismen van natuurlijke intelligentie, maar ook een bijdrage leveren aan het beter begrijpen van de processen die ten grondslag liggen aan verschillende hersenaandoeningen.



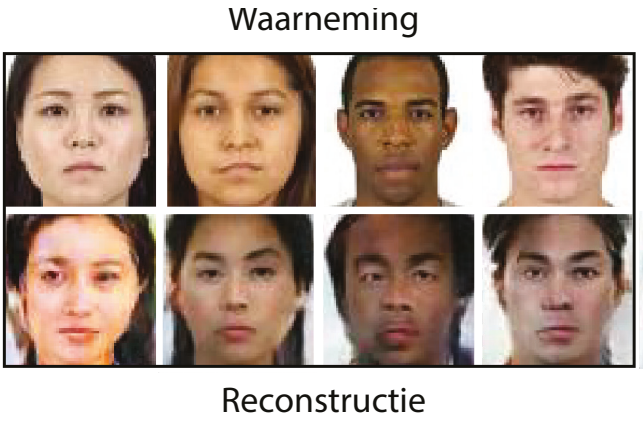
Figuur 16: Schatting van causale relaties tussen neurale populaties.

We zien dus dat intelligente technologie inzicht kan geven in onszelf. De fundamentele vraag vanuit theoretisch oogpunt blijft echter waarom er überhaupt rationele wezens bestaan die bepaalde doelen nastreven. De ultieme theorie zal ons dan ook moeten vertellen hoe natuurlijke intelligentie ontspringt vanuit primaire fysische processen. Wellicht is er een onderliggend principe dat ons vertelt waarom het brein zichzelf uitvindt onder evolutionair gereguleerde selectiedruk[56]. Een interessant voorbeeld van zo'n principe is *empowerment*. Dit principe formaliseert het idee dat organismen zichzelf dwingen om alle opties open te houden, zodat ze optimaal voorbereid zijn op ieder mogelijk toekomstscenario[57].

MENS-MACHINE INTERACTIE

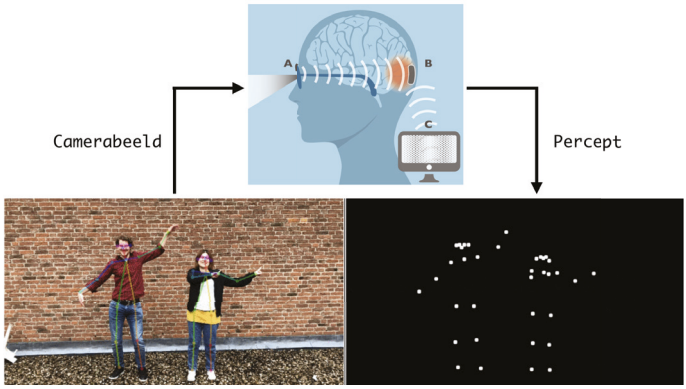
We hebben gezien dat de mens en de machine elkaar wederzijds kunnen inspireren. We kunnen echter ook een directe koppeling tussen mens en machine tot stand brengen. Door vooruitgang in de neurotechnologie kunnen we het brein steeds preciezer meten en stimuleren. Door neurotechnologie met AI te combineren kunnen we nieuwe vormen van mens-machine interactie tot stand brengen.

Aan de ene kant kunnen we neurale informatie in steeds meer detail uitlezen. Zo hebben wij computermodellen ontwikkeld die op basis van hersenactiviteit kunnen reconstrueren wat we hebben waargenomen[58] (figuur 17). Het steeds preciezer kunnen uitlezen van neurale activiteit geeft dus niet alleen inzicht in ons brein, maar ook in onze geest. Ooit kunnen we hiermee misschien verlamde mensen laten communiceren op basis van hun gedachten of toegang krijgen tot onze dromen[59].



Figuur 17: Het uitlezen van informatie in het brein. Boven ziet u vier gezichten zoals waargenomen door een proefpersoon in een MRI scanner. Onder ziet u de reconstructie van deze gezichten op basis van hersenactiviteit.

Aan de andere kant stelt de neurotechnologie ons in staat om het brein met steeds grotere precisie te beïnvloeden. Door neurale implantaten aan te sturen met behulp van AI kunnen we specifieke hersentoestanden opwekken. Binnen het Nestor consortium ontwikkelen wij intelligente technologie die op deze manier visuele sensaties tot stand kan brengen (figuur 18). Hiermee hopen we ooit blinde mensen een stukje visuele waarneming terug te kunnen geven[60]. Het herstellen of misschien zelfs wel verbeteren van hersenfuncties begint dus door middel van deze technologie binnen ons bereik te komen.



Figuur 18: Herstel van visuele waarneming door middel van AI en neurotechnologie.

ONTSPOORDE MACHINES

Ik heb u in deze voordracht verteld over hoe we menselijkere machines kunnen bouwen. Het idee van menselijke machines die ons voorbij kunnen streven is fascinerend, maar ook afschrikwekkend. De singulariteit, waarbij machines zo geavanceerd worden dat wij hen niet meer kunnen bijbenen, is dan ook een terugkerend thema in de science fiction. De bizarre situatie waartoe dit kan leiden wordt pijnlijk duidelijk in de film *Her*. Hierin wordt de hoofdpersoon verliefd op het intelligente besturingssysteem van zijn computer. Het besturingssysteem onderhoudt echter op zijn beurt simultaan relaties met duizenden mensen. Ze ontstijgt daardoor al snel de beperkte cognitieve capaciteiten van de mensheid als geheel. Ook het vervagen van de grens tussen mens en machine stemt tot overpeinzing. Zo laat de film *The Ghost in the Shell* zien hoe de menselijke geest los van het lichaam voort zou kunnen leven in een machine. Een andere vraag is hoe we menselijke machines tegen onszelf zouden moeten beschermen mocht de tijd daar zijn. Zo benadrukt Mary Shelley in haar boek hoe het monster van Frankenstein als creatie van de mens zelf wordt verstoten door zijn schepper[61]. U kunt zich afvragen hoe u om zou gaan met deze kunstmatige medeburgers.

Velen voor mij hebben op de gevaren van ontspoorde AI gewezen en het is belangrijk om stil te staan bij potentiële doemscenario's. We moeten echter ook realistisch zijn. Menselijke machines zijn dan wel een theoretische mogelijkheid, maar praktisch gezien zijn we hier nog ver van verwijderd. Onze huidige machines komen bij lange na niet in de buurt van de geavanceerde machinerie van het leven. Ook blijft onbeantwoord of menselijke machines noodzakelijkerwijs dezelfde subjectieve (fenomenologische) ervaringen zullen hebben als mensen. We noemen dit ook wel het moeilijke probleem van bewustzijn[62]. Dit probleem heeft de potentie om alsnog onze ideeën over denkende machines als een kaartenhuis ineen te laten storten.

Een veel urgenter gevaar is hoe de huidige (nog relatief stupide) intelligente technologie ingrijpt in de hedendaagse samenleving. AI kan namelijk ingezet worden om zowel onze waarneming als ons gedrag te monitoren, te manipuleren of zelfs volledig over te nemen. Er zijn helaas al voorbeelden te over. Denk bijvoorbeeld aan China's sociale kredietsysteem, waarin AI wordt gebruikt om de bevolking in de gaten te houden, de opkomst van apps die nepnieuws kunnen creëren dat niet van echt te onderscheiden is, of het debat over het gebruik van autonome wapens die zelf kunnen beslissen over leven en dood.

GETEMDE TECHNOLOGIE

Aan de andere kant biedt AI ons de middelen om beter voor onze planeet te zorgen en het welzijn van onze medemens te vergroten[63]. Intelligente technologie kan ons bijvoorbeeld helpen om bij te dragen aan het verwezenlijken van de doelstellingen op het gebied van duurzame ontwikkeling, zoals opgesteld door de Verenigde Naties[64] (figuur 19). De keuze is uiteindelijk aan ons[65]. Maar wat zijn dan de concrete stappen die we nú kunnen zetten?

Figuur 19: De doelstellingen van de VN op het gebied van duurzame ontwikkeling.

Ten eerste moeten wij als AI-onderzoekers misschien niet enkel *slimme* machines, maar vooral *wijze* machines bouwen die ons helpen om boven onze eigen beperkingen uit te stijgen. Dit betekent dat we moeten nadenken over hoe we machines kunnen voorzien van essentiële kwaliteiten, zoals nieuwsgierigheid, compassie, creativiteit en integriteit. Dit geeft een diepere invulling aan het idee van een menselijke machine.

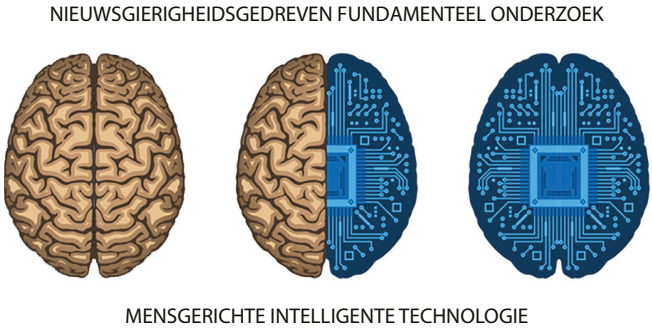
Ten tweede moeten we de AI die gebruikt wordt in intelligente toepassingen juist ook reguleren. Het is bijvoorbeeld nogal onzinnig als onze zelfrijdende auto onderweg aan het mijmeren is over de zin van het leven. We dienen transparante intelligente technologie te ontwikkelen die bewijst dat ze voldoet aan onze wensen. Dit laatste is echter niet triviaal. Want over wiens wensen hebben we het eigenlijk? In het geval van

de zelfrijdende auto doemt bijvoorbeeld de vraag op wanneer uw auto uw leven prijs zou moeten geven om andere weggebruikers te redden[66]. Hier komt de formalisering van rationaliteit weer om de hoek kijken: Welke utiliteitsfunctie moet de machine maximaliseren?

Ten derde dienen we het publiek en de overheid bewust te maken van de *mogelijkheden* en *beperkingen* van AI. Het is niet overdreven om te stellen dat de wetgevende, uitvoerende en rechterlijke macht sterk achterlopen als het gaat om de vraag onder welke condities we intelligente technologie zouden mogen en willen inzetten. Het antwoord op deze vragen vereist zowel een nauwe samenwerking tussen de alfa-, bèta- en gammawetenschappen, als een open dialoog tussen de academische wereld en de rest van de maatschappij.

DE TOEKOMST VAN AI

Wat betekent dit alles voor het AI-onderzoek aan de Radboud Universiteit? We zouden ons naar mijn mening moeten richten op twee speerpunten (figuur 20). Ten eerste moet nieuwsgierigheidsgedreven onderzoek naar de theoretische fundamenten van natuurlijke en artificiële intelligentie voorop staan. Dit vereist een multidisciplinaire aanpak waar neuropsychologische inzichten, wiskundige formaliseringen en fysische implementaties elkaar vinden. Ten tweede moeten we ons richten op de ontwikkeling van *mensgerichte* intelligente technologie, waarin het welzijn van de mens en zijn omgeving centraal staan[67]. Dit vereist dat we methodiek ontwikkelen om de maatschappelijke inzet van intelligente toepassingen op een verantwoorde manier te laten plaatsvinden. Het Donders Centrum voor Cognitie biedt de ideale omgeving voor dit multidisciplinaire onderzoek. Ooit ontstaan vanuit het Nijmegen Instituut voor Cognitie en Informatie heeft het altijd al het begrijpen van de mens in de machine en de machine in de mens als missie gehad. Tegelijkertijd zijn intensieve samenwerking met zowel andere wetenschappers als het bedrijfsleven onmisbaar om een leidende rol te blijven spelen op het gebied van de AI.



Figuur 20: Theoretische en toegepaste speerpunten van AI-onderzoek in Nijmegen. Nieuwsgierigheidsgedreven fundamenteel onderzoek naar natuurlijke en artificiële intelligentie en de ontwikkeling van mensgerichte intelligente technologie.

Ook in het AI-onderwijs staan de eerdergenoemde theoretische en toegepaste speerpunten hoog in het vaandel. Het is onze taak om onze studenten breed op te leiden, zodat ze niet alleen de *technische* vaardigheden hebben om de AI van de toekomst te bouwen, maar ook over de *academische* kwaliteiten beschikken om te reflecteren op de maatschappelijke impact van hun werk. Als academici hebben wij de taak om juist te vertragen in een almaar versnellende maatschappij, waarin te weinig wordt stilgestaan bij de consequenties van ons handelen.

Een voorwaarde voor excellentie in onderzoek en onderwijs is dat we een klimaat creëren, waarin stafleden de ruimte krijgen om vrij onderzoek te doen en studenten de aandacht krijgen die ze verdienen. Hier komt het begrip empowerment opnieuw om de hoek kijken. Volgens Maria Montessori, naamgeefster van het toekomstige gebouw van de Faculteit der Sociale Wetenschappen, hebben wij een natuurlijke drang tot zelfontplooiing[68]. De universiteit dient de juiste academische omgeving te bieden om deze zelfontplooiing mogelijk te maken. AI staat hier met de sterke groei van het aantal studenten, de dominante positie van techgiganten en de competitie met andere academische instellingen voor een grote uitdaging. Gezonde groei van AI zal gepaard moeten gaan met toenemende investeringen, zodat de kwaliteit van onderzoek en onderwijs gewaarborgd blijft. Ook op nationaal niveau zijn nieuwe investeringen en samenwerkingsverbanden cruciaal willen we als kenniseconomie niet drastisch achter gaan lopen op het internationale speelveld.

Ik heb het in deze voordracht gehad over menselijke machines. Het wordt dus tijd dat ik antwoord geef op de vraag of we menselijke machines kunnen bouwen. Het antwoord luidt vooralsnog nee. Er zijn nog steeds geen androïden die van elektrische schapen dromen[69]. Het grote wonder is echter dat, alhoewel we ze nog niet kunnen bouwen, ze al wel in ons midden zijn. Kijkt u maar eens naar de menselijke machines links of rechts van u. Hoe de geest kan ontstaan uit levenloze materie is voor mij de ultieme wetenschappelijke vraag. Er ligt een grote schoonheid besloten in de geest die zichzelf tot object van studie maakt. De wetenschap kan ons misschien antwoord geven op hoe het denken tot stand komt. Ze zal ons echter nooit kunnen vertellen hoe het *voelt* om een denkend wezen te zijn. Het is de kunst die ons toegang kan geven tot de subjectieve ervaring van bezielde materie[70]. De verwondering van de zichzelf beschouwende geest komt tot uitdrukking in de eerste strofe van het volgende gedicht van Emily Dickinson:

The Brain - is wider than the Sky -
For - put them side by side -
The one the other will contain
With ease - and You - beside -

DANKWOORD

Aan het eind gekomen van mijn oratie wil ik nog enkele woorden van dank uitspreken. Ten eerste wil ik het College van Bestuur, Michiel Kompier, decaan van de Faculteit der Sociale Wetenschappen, en het overige faculteitsbestuur danken voor het in mij gestelde vertrouwen. Ook dank ik Pieter Medendorp, directeur van het Donders Centrum voor Cognitie, en Ruud Meulenbroek, directeur van het Onderwijsinstituut voor Psychologie en Kunstmatige Intelligentie, voor het vertrouwen en de steun die ze mij hebben geboden tijdens het turbulente eerste jaar van mijn aanstelling.

Het bestuur en management van het Donders Instituut wil ik bedanken voor de mogelijkheden die ze mij geboden hebben om interdisciplinair onderzoek naar menselijke machines te verrichten over de grenzen van de faculteiten heen. Peter Desain wil ik danken voor de mogelijkheid die hij mij acht jaar geleden heeft geboden om als staflid bij AI te beginnen. Mijn collega's binnen het Donders Instituut wil ik bedanken voor de langdurige en vruchtbare samenwerking. Mijn collega's bij de Faculteit der Natuurwetenschappen, Wiskunde en Informatica wil ik bedanken voor het mogelijk maken van mijn promotieonderzoek op het gebied van de medische informatica en de samenwerking in zowel onderwijs en onderzoek. Specifiek wil ik Tom Heskes bedanken voor zijn betrokkenheid gedurende de afgelopen jaren.

Ook dank ik alle stafleden van de opleiding AI, het secretariaat AI, medewerkers van het Onderwijsinstituut en leden van het bestuur van de studievereniging CognAC. Ik realiseer me terdege dat we een hectische tijd achter de rug hebben. Zonder jullie toewijding en volharding was AI in Nijmegen al lang geleden opgehouden te bestaan. Tegelijkertijd kijk ik vol vertrouwen naar een mooie toekomst, waarin ieders unieke kwaliteiten op waarde worden geschat en we de ruimte krijgen om te excelleren in onderzoek en onderwijs. Ook dank ik alle studenten AI. Jullie vormen tezamen de ruggengraat van onze opleiding.

Ik heb u vandaag een deel van ons eigen onderzoek laten zien. Uiteindelijk zijn het de gedreven onderzoekers in de Artificial Cognitive Systems groep die veel van het werk hebben verricht. Umut, Yağmur, Max, Linda, Jan-Pieter, Luca, Sander, Elsbeth, Nadine, Silvan, Gabi, Katja en Jordy, het is een voorrecht om iedere dag weer samen met jullie onderzoek te mogen doen naar de aard van ons denken en menselijke machines een stapje dichterbij te brengen.

Bovenal dank ik mijn familie, schoonfamilie en vrienden. Lieve pa en ma, bedankt voor de onvoorwaardelijke steun en het grenzeloze vertrouwen dat jullie mij al die jaren hebben gegeven. Lieve Esther, als ik de machine ben, dan ben jij de mens. Als ik weer eens doorratel, dan ben jij het die me altijd weer laat terugkeren naar wat écht belangrijk is. Zonder jou had ik hier dan ook niet gestaan. Lieve Maya en Noor, wat is het geweldig om jullie te mogen zien opgroeien tot twee prachtige mensen. Ik plaag jullie weleens dat papa en mama jullie mooi geknutseld hebben, maar geloof me meiden, dáár kunnen de machines nog wat van leren.

Ik heb gezegd.

NOTEN

1 Ik negeer hier voor het gemak het waarnemen van lichamelijke sensaties en het produceren van reacties die invloed op onze interne toestand uitoefenen.

2 Zie R. Descartes, *Principia Philosophiae*, 1644. Tekening van René Descartes in Treatise of Man. De gesloten kring die hiermee ontstaat, noemen we ook wel de perceptie-actie cyclus. Zie J. M. Fuster, Upper processing stages of the perception-action cycle. *Trends in Cognitive Sciences*, 8(4):143–145, 2004.

3 S. Herculano-Houzel, The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3: 31, 2009.

4 Dit wordt ook wel mentale chronometrie genoemd. Zie F. C. Donders, On the speed of mental processes. *Acta Psychologica* 30: 412-31, 1969.

5 Foto's afkomstig uit de collectie Universiteitsmuseum Utrecht.

6 N. Dijkstra, S. Bosch en M. A. J. van Gerven. Vividness of visual imagery depends on the neural overlap with perception in visual areas. *The Journal of Neuroscience*. 37(5): 1367-1373, 2017.

7 Zie bijvoorbeeld T. Hobbes, *Leviathan, or The Matter, Forme & Power of a Common-Wealth Ecclesiasticall and Civil*, 1651.

8 A. M. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, 42(1): 230–65, 1936.

9 Zo'n honderd jaar eerder kwam Charles Babbage al met een concept voor een programmeerbare rekenmachine genaamd de Analytical Engine. Zijn protégé, Lady Ada Lovelace wordt alom gezien als de eerste programmeur. De moderne computer is ook schatplichtig aan andere briljante onderzoekers zoals Kurt Gödel en Konrad Zuse. Zie J. Schmidhuber, Turing in Context, Science, 336(6089): 1638-1639, 2012.

10 Alan Turing by Elliott & Fry. Quarter-plate glass negative, 29 March 1951, NPG x82217 © National Portrait Gallery, London.

11 Zowel de AI als de cognitiewetenschap vinden hun oorsprong rondom de Dartmouth workshop, die in 1956 werd georganiseerd (https://en.wikipedia.org/wiki/Dartmouth_workshop).

12 Met menselijke machines bedoel ik dus hetzelfde als strong AI of Artificial General Intelligence. Om te testen of we machines nog van mensen kunnen onderscheiden, stelde Alan Turing de Turing test voor. Zie A. M. Turing, Computing machinery and intelligence. *Mind*, 49, 433–460.

13 Actie wordt ook binnen de cognitieve psychologie als primair gezien. Zoals verwoord door Roger Sperry: *The entire output of our thinking machine consists of nothing but patterns of motor coordination*. Zie R. W. Sperry, Neurology and the mind-brain problem. *American Scientist*, 40(2): 291–312, 1952.

14 Ook utiliteit staat centraal in ons denken over natuurlijke intelligentie. Zo stelde John Dewey al in 1896: *There is simply a continuously ordered sequence of acts, all adapted in themselves and in the order of their sequence, to reach a certain objective end, the reproduction of the species, the preservation of life, locomotion to a certain place*. Zie J. Dewey, The reflex arc concept in psychology. *Psychological Review*, 3: 357–370, 1896.

15 Inherente onzekerheid door de stochastische aard van onze wereld noemen we ook wel *aleatoric uncertainty*. Systematische onzekerheid door kennis die we niet hebben maar wel zouden kunnen hebben, noemen we *epistemic uncertainty*.

16 We noemen dit het maximum expected utility principe. Zie J. von Neumann en O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

17 T. Bayes, An essay towards solving a problem in the doctrine of chances. Herdruk in *Philosophical Transactions*, 53: 370-418, 1983.

18 Zie bijvoorbeeld C. P. Robert, The Bayesian Choice: A Decision-Theoretic Motivation, Springer, 1994 en E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003 voor een uiteenzetting van het Bayesiaanse gedachtengoed.

19 Zie D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003 voor een beschrijving van stochastische en deterministische inferentiemethoden. Ook in ons eigen werk hebben wij een aantal van deze methoden ontwikkeld. Zie M. Hinne, A. Lenkoski, T. Heskes en M. A. J. van Gerven, Efficient sampling of Gaussian graphical models using conditional Bayes factors. Stat, (0373), 2014 en L. Ambrogioni, U. Güçlü, Y, Güçlütürk, M. Hinne, E. Maris en M. A. J. van Gerven, Wasserstein variational inference, *Advances in Neural Information Processing Systems*, 2018.

20 M. A. J. van Gerven, B. G. Taal en P. J. F. Lucas, Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics*, 41(4): 515–529, 2008.

21 Zie bijvoorbeeld K. Doya, S. Ishii, A. Pouget en R. P. N. Rao, Bayesian Brain, Cambridge, MA: The MIT Press, 2007. Beschrijvingen van ons brein in deze termen vinden hun oorsprong in het werk van onder andere Alhazen en Helmholtz. Zie A. I. *Sabra, ed., The Optics of Ibn al-Haytham, Books I–II–III: On Direct Vision*, London: The Warburg Institute, University of London, 1989 en H. von Helmholtz, *Handbuch der Physiologischen Optik*, Leipzig: Voss, 1867. Deviaties van optimaliteit zijn te verklaren door te appelleren aan de bounded resources die een organisme tot zijn beschikking heeft. Zie H.A. Simon, *Models of Bounded Rationality*, vols. 1+2, The MIT Press, Cambridge, MA, 1982.

22 Zie R. Grush, The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3): 377–396, 2004.

23 R. Gregory, *The Intelligent Eye*, Weidenfeld and Nicolson, 1970.

24 Zoals bijvoorbeeld de cybernetica, Zie N. Wiener, *Cybernetics, Or Control and Communication in the Animal and the Machine* (2nd ed.), The MIT Press, 1961. Zie ook G. Dyson, *Darwin Among the Machines: The Evolution of Global Intelligence*, Basic Books, 2012.

25 Een klassiek voorbeeld is het Hodgkin en Huxley model dat het genereren van actiepotentialen door individuele neuronen beschrijft. Zie A. L. Hodgkin en A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4): 500, 1952.

26 Deze tak van onderzoek komt voort uit het werk van Warren McCullloch en Walter Pitts, die aantoonden dat eenvoudige neurale netwerken gebaseerd op logische elementen dezelfde expressiviteit hebben als universele Turingmachines. Zie W. S. McCulloch en W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4): 115–133. 1943.

27 Een van de eerste leerregels is de perceptron leerregel. Zie F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review*, 65: 386-408, 1958. Backpropagation (of error) is gebaseerd op het creatief gebruiken van de kettingregel zoals gehanteerd binnen de differentiaalrekening. Ze is sterk gerelateerd aan het Gauss-Newton algoritme en is veelvuldig opnieuw uitgevonden. Zie P. J. Werbos, *The Roots of Backpropagation. From Ordered Derivatives to Neural Networks and Political Forecasting*. John Wiley & Sons, 1994.

28 Het connectionisme won sterk aan populariteit in de jaren tachtig van de vorige eeuw door de opkomst van de Parallel Distributed Processing beweging. Zie McClelland, J., Rumelhart, D., *et al*., 1986, *Parallel Distributed Processing*, Volume I en II, The MIT Press.

29 We kunnen parallele informatieverwerking wel *emuleren* op Von Neumann-architecturen.

30 Y. LeCun, Y. Bengio en G. Hinton, Deep learning. *Nature, 521*(7553): 436–444, 2015.

31 Eerder veronderstelde men dat diepe neurale netwerken niet te trainen waren. Echter, door een aantal wiskundige verbeteringen, de beschikbaarheid van steeds meer data en steeds snellere hardware kwam het leren van deze modellen binnen handbereik. Het kwam erop neer dat neurale netwerken langer getraind moesten worden om realistische problemen op te lossen.

32 Y. Güçlütürk, U. Güçlü, R. van Lier, en M. A. J. van Gerven, Convolutional sketch inversion. *Lecture Notes in Computer Science*, Vol. 9913 LNCS, 2016.

33 De eerste moderne computers werden ontwikkeld in de Tweede Wereldoorlog. Zo ontwikkelden de Britse geallieerden op Bletchley Park de Colossus om geheime codes van de Nazi's te breken. In de Verenigde Staten werd de ENIAC ingezet om berekeningen uit te voeren bij de ontwikkeling van de eerste atoombom.

34 Het spel Go is een googol (10100) keer complexer dan schaken De schaakcomputer Deep Blue versloeg schaakgrootmeester Gary Kasparov in 1997, maar maakte hierbij nog uitgebreid gebruik van menselijke expertise. Zie: G. Kasparov, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. PublicAffairs, 2017.

35 D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang et al., Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359, 2017.

36 https://blog.openai.com/openai-five

37 Zie V. Fischer, M. C. Kumar, J. H. Metzen en T. Brox, Adversarial examples for semantic image segmentation. *In International Conference on Learning Representations*, 1-4, 2017.

38 Het niet kunnen generaliseren van simulaties naar de realiteit noemt men ook wel de *reality gap*.

39 J. R. Searle, Minds, brains and programs. *Behavioral and Brain Sciences*, 3: 417-457, 1980.

40 We noemen dit *active sensing*. Zie S. C. H. Yang, D. Wolpert en M. Lengyel, Theoretical perspectives on active sensing. *Current Opinion in Behavioral Sciences, 11*: 100–108, 2016.

41 Ook de zelflerende capaciteiten van synthetische breinen zouden we kunnen verbeteren door inzichten over biologische leerprocessen op verschillende tijdschalen beter te integreren. Zie bijvoorbeeld E. Bates, J. Elman, M. H. Johnson, A. Karmiloff-Smith, D. Parisi en K. Plunkett, *Rethinking Innateness*. The MIT Press, 1996. Ook de biofysica vertelt ons hoe we realistischere modellen kunnen ontwikkelen. Zie bijvoorbeeld W. Bialek, *Biophysics: Searching for Principles*. Princeton University Press, 2012.

42 Modeling and Simulating Eye Oculomotor Behavior to Support Retina Implant Development. Auteurs: Ra'anan Gefen en Leonid Yanovitz, Nano-Retina, Inc.

43 D. H. Hubel en T. N. Wiesel, Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148: 574–591, 1959.

44 Zie bijvoorbeeld J. Guergiuev, T. Lillicrap en B. A. Richards, Towards deep learning with segregated dendrites, *arXiv:1610.00161v3*: 1–4,1, 2017 en J. H. Lee, T. Delbruck en M. Pfeiffer, Training deep spiking neural networks using backpropagation, *ArXiv:1608.08782:* 1–10, 2016.

45 Hierdoor zijn we misschien ooit in staat om intelligentie direct in de structuur van de materie te verweven Zoals Feynman al zei: *There is plenty of room at the bottom*.

46 Volgens David Marr en Tomaso Poggio kunnen we een informatieverwerkingssysteem beschouwen op drie verschillende verklaringsniveaus . Op het computationele niveau vragen we ons af welk probleem het systeem oplost en waarom. Op het algoritmische niveau vragen we ons af op welke wijze het systeem dat probleem oplost. Op het implementatie niveau vragen we ons af hoe deze oplossing fysisch gerealiseerd is. Zie D. Marr en T. Poggio, *From understanding computation to understanding neural circuitry*, A.I. Memo, 1-22, 1976. Poggio voegt hier later nog het leren op het niveau van het individu en de soort aan toe om te verklaren hoe het probleemoplossend vermogen van een informatieverwerkingssysteem tot stand komt. Zie T. Poggio, *The levels of understanding framework, revised*. MIT-CSAIL-TR-2012-014 CBCL-308, 1-11, 2012.

47 Predictive processing. A. Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2016. Zie ook K. J. Friston, The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, 11(2): 127–138, 2010.

48 https://deepmind.com/blog/neural-scene-representation-and-rendering

49 Dit sluit aan op de emulatietheorie die stelt dat het denken gelijk is aan gesimuleerd gedrag. Zie R. Grush, The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3): 377–396, 2004. Zoals Alexander Bain stelde: *Thinking is restrained speaking or acting*. Zie A. Bain, *The Senses and the Intellect*, 1894.

50 Zoals geopperd in C. M. A. Pennartz, *The Brain's Representational Power*, The MIT Press, 2015.

51 U. Güçlü en M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, 35(27): 10005–10014, 2015.

52 Dit komt neer op de acquisitie van 120.000 FMRI scans. Dit is zo mogelijk de langste naturalistische dataset ooit verzameld in één proefpersoon.

53 M. Hinne, A. Meijers, R. Bakker, P. H. E. Tiesinga, M. Mørup en M. A. J. van Gerven, The missing link: Predicting connectomes from noisy and partially observed tract tracing data. *PLoS Computational Biology*, 13(1): e1005374, 2017.

54 Zie R. J. Janssen, P. Jylänki en M. A. J. van Gerven, Let's not waste time: Using temporal information in clustered activity estimation with spatial adjacency restrictions (CAESAR) for parcellating FMRI data. PLoS ONE, 11(12), 1–21, 2016 en M. Hinne, R. J. Janssen, T. Heskes en M. A. J. van Gerven, Bayesian estimation of conditional independence graphs improves functional connectivity estimates. *PLoS Computional Biology,* 11(11): e1004534. 2015.

55 L. Ambrogioni, P. Ebel, M. Hinne, U. Güçlü, M. A. J. van Gerven en E. Maris, Semi-analytic nonparametric Bayesian inference for spike-spike neuronal connectivity. *BioRxiv,* 340489, 2018.

56 Onderzoek naar artificieel leven kan belangrijke inzichten bieden. Zie bijvoorbeeld S. Kauffman, *At Home in the Universe*, Oxford University Press, 1996 en M. A. Bedau, Artificial life: Organization, adaptation and complexity from the bottom up. *Trends in Cognitive Sciences*, 7(11), 505–512, 2003. Ook informatietheorie en statistische fysica bieden belangrijke aanknopingspunten. Zie C. E. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, 27: 379–423 & 623–656, 1948 en F. Seoane en R. V. Sol, Information theory, predictability, and the emergence of complex life, *arXiv:1701.02389v2*: 1–13, 2017.

57 C. Salge, C. Glackin en D. Polani, Empowerment - an introduction. ArXiv:1310.1863: 1–46. 2013

58 Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier en M. A. J. van Gerven, Deep adversarial neural decoding. *Advances in Neural Information Processing Systems*, 1–12, 2017.

59 Zie J.-D. Bauby, *The Diving Bell and the Butterfly*, Vintage, 1998 voor een beschrijving van hoe het is om niet meer te kunnen communiceren. We noemen dit het locked-in syndrome.

60 Zie https://nestor-sight.com. Zie het boek J. Naumann, *Search for Paradise*, Xlibris, 2012 voor een eerstehands beschrijving van hoe het is om weer te kunnen zien met behulp van een corticaal implantaat.

61 M. W. Shelley, *Frankenstein, Or, The Modern Prometheus*: the 1818 Text. Oxford University Press, 1998.

62 D. Chalmers, Facing up to the problem of consciousness, *Journal of Consciousness Studies*, 2(3): 200-219.

63 Zie bijvoorbeeld Y. N. Harari, Homo Deus: A Brief History of Tomorrow, Vintage Books, 2017.

64 https://sustainabledevelopment.un.org

65 Zoals verwoord door Stephen Hawking: *Success in creating effective AI, could be the biggest event in the history of our civilization. Or the worst. We just don't know. So we cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it.*

66 Dit is een variant van het Trolley problem. Zie P. Foot, The problem of abortion and the doctrine of the double effect. In: *Virtues and Vices*, Basil Blackwell, 1978.

67 Zie ook H. G. Rickover, A Humanistic Technology. *Nature*, 5012: 721–726, 1965.

68 M. Montessori, *The Montessori Method*. Frederick Stokes, 1912.

69 P. K. Dick, *Do Androids Dream of Electric Sheep?* Ballantine Books, 1968.

70 Zie J. Lehrer, *Proust was a Neuroscientist*, Canongate Books, 2012 voor een uiteenzetting van de relatie tussen kunst en wetenschap. Zie ook T. Nagel, What is it like to be a bat? *Philosophical Review*, 83(4): 435-450, 1974 voor een verkenning van objectieve versus subjectieve verklaringen van het mind-body probleem.